Final Project - DA2 & Coding 1 - Aftab Alam

Introduction

The objective of this paper is to explore the economic concept of price elasticity of demand (PED) for a few products listed on the eCommerce website **Wish.com**. The intention is to see how, on average, customers responded to a change in price while controlling for a few factors, where that product is the only listed product by the merchant on their page.

$$PED = \%\ Change\ in\ Quantity\ /\ \%\ Change\ in\ Price$$

The paper will use the following regression formula to estimate the average PED for the products that is the only one listed by the merchant, where 'Z' refers to the confounding variables.

$$log(Quantity\ Sold)\ = \beta_0 + \beta_1\ log(Online\ Prices)\ + Z$$

It would be difficult to hypothesize before doing any analysis, whether the PED would be elastic or inelastic for the products used in this paper as the product specifications are not available and the only information available is whether a certain merchant lists only 1 product on their page or any other number of products.

The paper uses a number of techniques to analyze the above said concept that includes running simple linear regression, non-parametric LOWESS regressions, and parametric linear regressions.

Data

The data itself has been taken from **Data.World**. It was posted by a user who was looking at the best product in terms of sale and had prepared a cleaned version, which has been utilized in the paper. However, contrary to the analysis conducted by that user, this paper is looking to see how elastic the demand is for those products for a percentage change in price. Here, for quantity demanded, products sold is taken as a proxy variable.

The data set originally contained around 900 observations for unique merchants who had listed their products on the Wish.com and the data was obtained as of July, 2020. It contains variables related to the merchant, product prices, and ratings. The paper focuses on the prices and total sales for merchants that have only listed 1 product.

Data Limitations The 'listedproducts' should not be confused with product code. It primarily gives the number of products listed by the seller on the website. For example, when the listed products is 1 for a few merchants, it does not necessarily mean they all list the same product, it just means they all list 1 product. Due to this, the paper is not looking for the price elasticity of demand for 1 particular product but due to this limitation, the paper is looking at price elasticity of demand for a single product listed by the merchants on the website.

The implication of this limitation is that the calculated price elasticity of demanded cannot be compared as the products might not be similar across merchants. This also means that the paper cannot associate the PED to a certain product. It, however, will try to give an average PED for all the different products as an average.

Nevertheless, the paper hypothesizes that the PED will be significantly less than 0 for a single product listed by the merchants on the website as per the law of demand, which states that quantity demanded decreases if the price of the product increases while keeping everything else constant. This is strictly about the sign of association on average and not the absolute magnitude.

$$H_0\ :\ \beta\ >=\ 0\ ,\quad H_a\ :\ \beta_1\ <\ 0$$

Important Variables

- 'listedproducts' is the number of products listed by the seller on the website.
- 'totalunitssold' is the total number of units sold for the selected product.
- 'rating' is the mean rating for the seller on the website.
- 'meanproductprices' is the mean price of the product online.
- 'meanretailprices' is the mean retail price of the product in brick & morter stores.
- 'meandiscount' is the mean discount provided to the customers on this product. This discount is applied on the retail price, which then gives us the mean online price.
- 'totalurgencycount' is whether there was an urgency banner displayed on the product page or not.

Data Cleaning & Munging The data taken from the Data.World was already clean and ready to use. However, I did make some changes to a few variables before using them in the regression analysis, based on a quick summary of the data set. Data summary of the base, unfiltered, data set is pasted in the appendix as 'Table 3'.

The original data set contained 13 variables in total, however, not all of these have been utilized.

- The variable 'averagediscount' was dropped from the dataset as it was duplicate of 'meandiscount' column
- The variable 'urgencytextrate' was dropped because it was a duplicate of 'totalurgencycount'; urgencytextrate = 100 * totalurgencycount
- Replaced the null values in the variable 'totalurgencycount' with 0 as null meant that there was no urgency banner associated to the product on the product page.
- The variable 'listedproducts' contained 11 unique values that referred to the number of products listed by each merchant on their online page. The variable contained the most observations (635 out of 958) for merchants who listed only 1 product on their page. So, for the purpose of this paper, I went ahead and selected the value of '1' for the number of products listed by merchants on their page.
- After making the above decision, the variable 'meanunitssoldperproduct' was then dropped because values of 'meanproductssold' and 'totalunitssold' were now identical, as the total units sold was divided by number of products to get the mean units sold of each list product, 1 in our case.

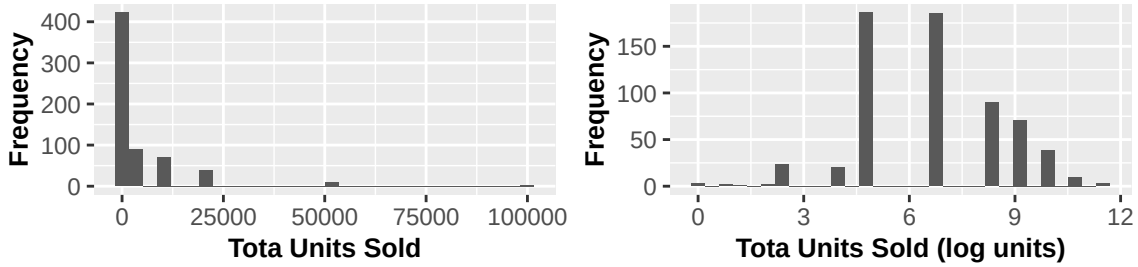
After making above decisions regarding the variables, the following table gives us a quick summary for the important variables we went ahead with to use in our analysis.

Table 1: Descriptive statistics

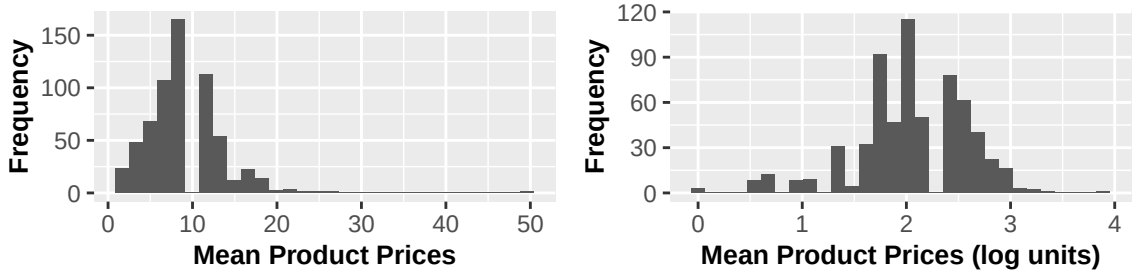
	Mean	Median	SD	Min	Max	P25	P75	N
Total Units Sold	4638	1000	10 240	1	100 000	100	5000	635
Rating	4	4	0	2	5	4	4	635
No. of Merchant Ratings	19 383	5048	93 200	0	2 174 765	1286	14 990	635
Mean Price	9	8	4	1	49	6	11	635
Mean Retail Price	25	11	31	1	252	7	30	635
Mean Discount	30	16	41	−18	97	−11	74	635
Product Rating Count	982	205	2127	0	20 744	32	920	635
Urgency Banner	0	0	0	0	1	0	0	635

The next step after identifying the important variables for our analysis was to look at how the individual variables are distributed and whether these need any transformations.

- The variable 'totalunitssold' had a skewed distribution with a long right tail. Hence, I created a new variable 'lnsales' that took the log of 'totalunitssold'. Although, the new distribution is not normal, I went ahead with it as it is not skewed.



- The variable 'meanproductprices' had a skewed distribution with a long right tail. Hence, I created a new variable 'ln_online_price' that took the log of 'meanproductprices' resulting in a near normal distribution curve.



- The variable 'rating' was distributed normally so it was taken as is. Graphs for the rest of these variables are placed in the appendix.
- The variable 'merchantratingscount' had a skewed distribution with a long right tail. Hence, I created a new variable 'ln_m_rating_count' that took the log of 'merchantratingscount' resulting in a near normal distribution curve.
- The variable 'meanretailprices' had a skewed distribution with a long right tail. Hence, I created a new variable 'ln_retail_price' that took the log of 'meanretailprices' resulting in a near normal distribution curve.
- The variable 'meanproductratingscount' had a skewed distribution with a long right tail. Hence, I created a new variable 'ln_p_rating_count' that took the log of 'meanproductratingscount' resulting in a near normal distribution curve.
- I don't have to look at the distribution of 'totalurgencycoun' as it is a binary variable.
- I did not look at the distribution of 'meandiscount' as it contains -ve values as well so taking its log will not be possible. Negative values suggest that the mean online price is higher than the mean retail price in the brick & morter shops, whereas a positive value suggests that the mean online price is lower than the mean retail price in the brick & morter shops.

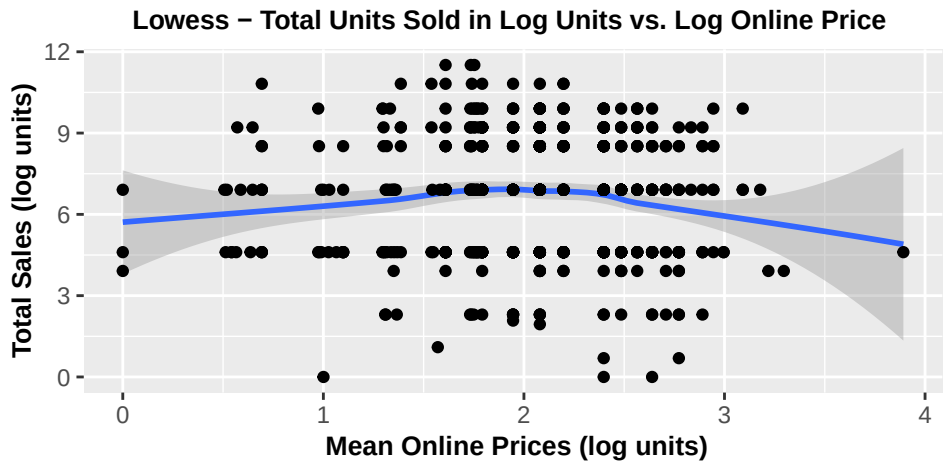
Analysis

Choosing Confounding Variables After all the data transformation and cleaning, I then looked at the possible correlations between the important variables discussed above. For this, I created a correlation matrix, shown in the appendix.

The dependent variable 'lnsales' has the weakest, near zero, correlation with 'ln_p_rating_count' and 'ln_m_rating_count' due to which these will not be used in the regression analysis ahead. Apart from these two variables, the rest have some kind of correlation that will be explored in the analysis ahead.

The near zero correlation of 'lnsales' with 'ln_p_rating_count' and 'ln_m_rating_count' could be explained based on the reason that maybe customers don't usually look at the total number of ratings provided for the product and/or the merchant and rather only focus on the mean rating when purchasing a product.

Regression Analysis Before moving on to estimate a regression model with confounding variables, the paper first looks at how the main y 'lnsales' and x 'ln_online_prices' are associated; whether this relationship is linear or not. To do this, I have first run a non-parametric LOWESS regression that results in the following graph.



Based on this LOWESS, it is visible that the association between these two variables might not be linear, hence, I might use splines in the parametric regression ahead. However, before directly jumping into using splines for the variable 'ln_online_prices', I ran a simple linear regression and a parametric splined regression to compare the beta coefficients and answer the question; are the beta coefficients in the splined linear regression significantly equal to the beta coefficient in the simple linear regression?

Upon running the regressions (Model 1: Simple Linear Regression vs Model 2: Linear Regression in the *Regression Models* table below) and comparing the confidence intervals (CI), it was difficult to conclude whether the beta coefficients from two regression models were significantly same or not as their CI overlapped but the model 1 beta values was outside of model 2 beta CI. A better way forward would be to formally test whether the betas are different or not, however, since the overlap of CIs is very small, for the purpose of this paper, I would assume the the Beta coefficients in two models are significantly different. More details on these CI for each of these beta coefficient are in the appendix.

After deciding on how to go about using the x variable 'ln_online_price', I looked at the association between 'lnsales' vs. 'rating', 'lnsales' vs. 'ln_retail_price', and 'ln_sales' vs. 'meandiscount' using non-paramteric LOWESS regressions. The graphs for these regressions are placed in the appendix.

- The association between 'lnsales' and 'rating' is non-linear hence I will be using splines for rating with a knot at 'rating' = 4.1

- The association between 'lnsales' and 'ln_retail_price' is non-linear hence I will be using splines for 'ln_retail_price' with knots at 'ln_retail_price' = 1.5 and 'ln_retail_price' = 3.5
- The association between 'lnsales' and 'meandiscount' is non-linear hence I will be using splines for 'meandiscount' with knots at 'meandiscount' = 15 and 'meandiscount' = 69
- I will be using the variable 'totalurgencycount' without any splines as this a binary variable

After identifying the association of above mentioned variables using the LOWESS curves, I have then added each variable incrementally one by one into linear regression models to see how each additional confounding variable impacts the beta coefficients of our main x variable 'ln_online_price'.

As summarized in the table 2 in the appendix below, when the confounding variable 'log(rating)' is added in model 3, it decreased the beta coefficients for both splines of the 'ln_online_price', estimating, in absolute terms, that on average the PED could be lower in the population, i.e. it could be inelastic.

Adding the next confounding variable 'meandiscount' into model 4 impacts the the beta coefficients of 'ln_online_prices' differently. In absolute terms, when the online price is lower than 2 log units, the PED is now higher compared to the previous model, makig it less inelastic. Whereas, in absolute terms, when the online price is greater than or equal to 2 log units, the PED is now lower compared to the previous model, making PED more inelastic.

Further, adding the 'totalurgencycount' confounding variable does not impact the coefficients of our main x variable 'ln_online_price' much and its own coefficient is not significant.

For model 6, adding the 'ln_retail_price' changes a lot of things in the regression model. In absolute terms, the coefficient for online price with the value of less than log unit 2, PED became more elastic compared to the previous model. Whereas, for online prices greater than or equal to 2 log units, in absolute terms, PED became more inelastic, moving nearer towards 0. Additionally, these coefficients are not significant at 95% confidence level. Some other coefficients that were previously significant also became non-significant.

Moreover, looking back at the data again and it's data summary, it turns out that ln_retail_price, meandiscount, and ln_online_price are highly correlated due to the fact that online prices are discounted compared to the retail prices and hence there is no need for inclusion of both mean discount and mean retail prices. Also looking back at the correlation matrix, it shows that retail price and discount are highly correlated.

Hence I will go forward with just keeping the discount variable and not include the retail prices in the model. I will also not include the 'totalurgencycount' variable as it does not seem to impact the beta of "ln_online_prices".

Lastly, comparing the BIC (Bayesian Information Criteria), it is higher for model 6, however, without the 'totalurgency' and 'ln_retail_price' variables in model 4, the BIC there is a lot lower even with confounding variables.

Conclusion

The preferred model for this paper is shown below using the parametric equation with the final beta coefficient values. The counfounding variables have been clubbed into the 'Z' variable.

$$log(Quantity\ Sold) \ = \ -12.28 + 0.81\ log(Online\ Prices)(< 2) \ - 0.85\ log(Online\ Prices)(>= 2) \ + Z$$

The model suggests that for online prices less than log 2 units (exp^2 = Euros 7.4), the PED is positive on average, suggesting that these products maybe be considered **Giffen** goods that violate the law of demand. Meaning a percentage higher price results on average in a percentage higher demand for the product on average. On the other hand, online prices greater than log 2 units (exp^2 = Euros 7.4), the PED is negative, as expected, hence, holding the law of demand true for this data set on average. Estimating a percentage higher price results on a average in a percentage lower demand for the product.

However, can we generalize these results to all the products available on the Wish.com? Possibly not. Because our data set has numerous limitations, including the fact that we do not have the details of each product. A better way forward would be to get prices of chosen products from Wish.com, collect their retail prices, and avail as many confounding variables as possible, such as country, size, color etc, so that based on the needs, those can be included in the regression model.

Why go this extra mile? This could possibly help merchants price their products better after learning how consumers react to price changes for specific products, hence managing their overall sales.

Appendix

[H]

Table 2: Regression Models to Explore Price Elasticity of Demand for Single Listed Product by a Merchant

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Intercept	6.81***	5.65***	−12.54***	−12.28***	−12.10***	−11.57***
	(0.35)	(0.46)	(2.53)	(2.50)	(2.51)	(2.52)
ln_online_price	−0.08					
	(0.17)					
Log of Online Prices (<2)		0.68***	0.65**	0.81***	0.78***	1.52*
		(0.26)	(0.25)	(0.27)	(0.27)	(0.89)
Log of Online Prices (>=2)		−1.01***	−0.89***	−0.85***	−0.85***	−0.30
		(0.32)	(0.32)	(0.32)	(0.32)	(0.94)
lspline(log(rating), 1.4)1			13.28***	13.07***	12.94***	12.89***
			(1.82)	(1.81)	(1.82)	(1.79)
lspline(log(rating), 1.4)2			−4.42	−3.14	−3.09	−3.50
			(3.47)	(3.49)	(3.49)	(3.50)
Mean Discount (<15)				0.02*	0.02*	0.02*
				(0.01)	(0.01)	(0.01)
Mean Discount (>=15 & <69)				−0.02***	−0.02***	−0.01
				(0.01)	(0.01)	(0.02)
Mean Discount (>=69)				0.03*	0.03*	0.06
				(0.02)	(0.02)	(0.06)
Urgency Banner?					0.20	0.19
					(0.19)	(0.19)
Log of Retail Prices (<2)						−0.99
						(0.96)
Log of Retail Prices (>=2 & <4)						−0.63
						(0.94)
Log of Retail Prices (>=4)						−0.07
						(0.99)
Num.Obs.	635	635	635	635	635	635
BIC	2819.3	2815.2	2783.7	2792.4	2797.9	2815.6

Regression results are calculated based on heteroskedastic robust standard errors
* p < 0.1, ** p < 0.05, *** p < 0.01

Table 3: Descriptive Statistics of the Raw Data

	Mean	Median	SD	Min	Max	N
No. of Products	2	1	1	1	15	958
Total Units Sold	7124	1000	14 364	1	120 000	958
Mean Units Sold	4408	1000	9167	1	100 000	958
Rating	4	4	0	2	5	958
No. of Merchant Ratings	22 020	5990	84 732	0	2 174 765	958
Mean Price	9	8	4	1	49	958
Mean Retail Price	25	11	30	1	252	958
Mean Discount	29	16	40	−18	97	958
Average Discount	29	16	40	−18	97	958
Product Rating Count	923	210	1926	0	20 744	958
Urgency Banner	1	1	1	1	6	391
Urgency Banner Rate	66	50	28	14	100	391

Details on beta coefficients of model 1 and model 2

CI of beta coefficient from model 1 for 'ln_online_price'

$$CI \ : \ [\ -0.41 \ , \ 0.25 \ ] \ where \ \beta \ = \ -0.08$$

The beta coefficient here estimates that if the online price is higher for the product by 1%, the average number of units sold will be lower by 0.08%, where it is not significant even at a 90% confidence interval. The CI suggests that the true beta on average in the population will be between -0.45 & 0.25, implying that we fail to reject our original null hypothesis that the PED in the population will be greater than or equal to zero for a single product listed by the merchants on the website.

CI of beta coefficient from model 2 for the first spline of 'ln_online_price'

$$CI \ : \ [\ 0.15 \ , \ 1.20 \ ] \ where \ \beta_1 \ = \ 0.68$$

Here the beta coefficient is significant for the first spline of 'ln_online_price' estimating that on average the PED is between 0.15 and 1.2 when the price of the product on the website is less than 2 log units, i.e around 7.4 Euros and we can reject the original null hypothesis with 99% confidence, i.e. on average PED is less than 0 in the population, holding law of demand true keeping everything else constant.

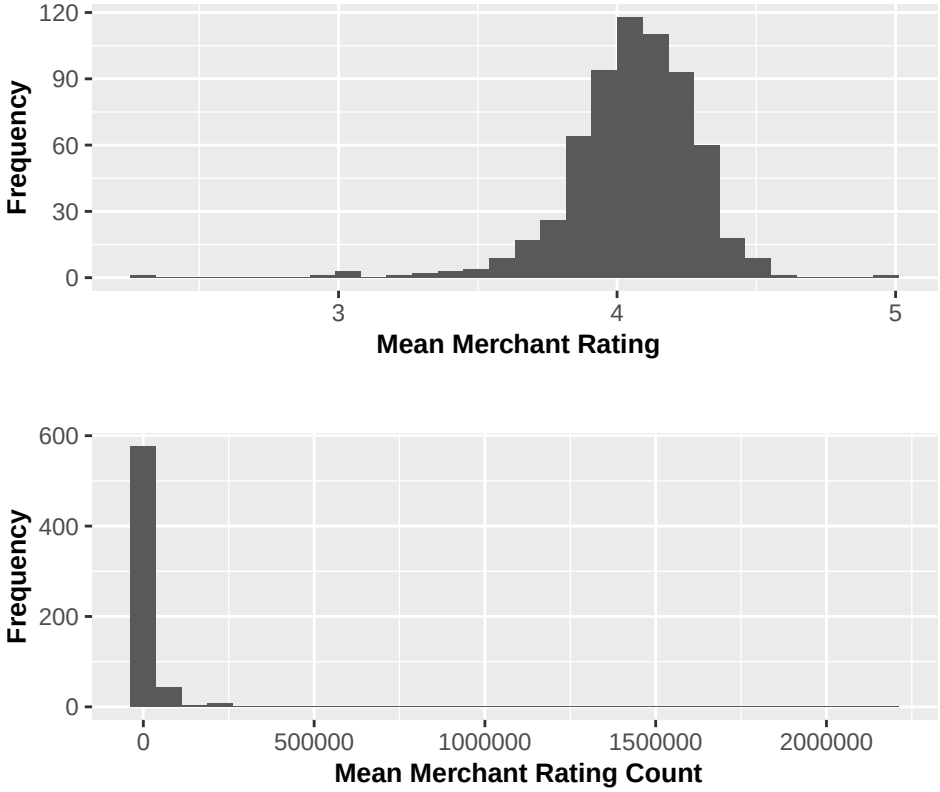
CI of beta coefficient from model 2

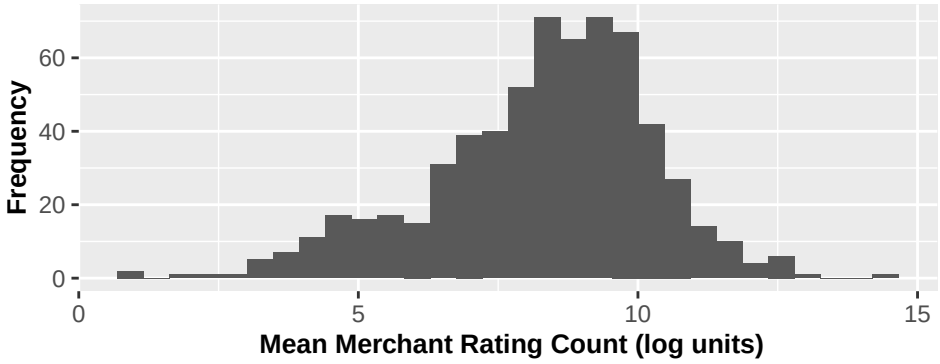
$$CI \ : \ [\ -1.66 \ , \ -0.37 \ ] \ where \ \beta_2 \ = \ -1.01$$

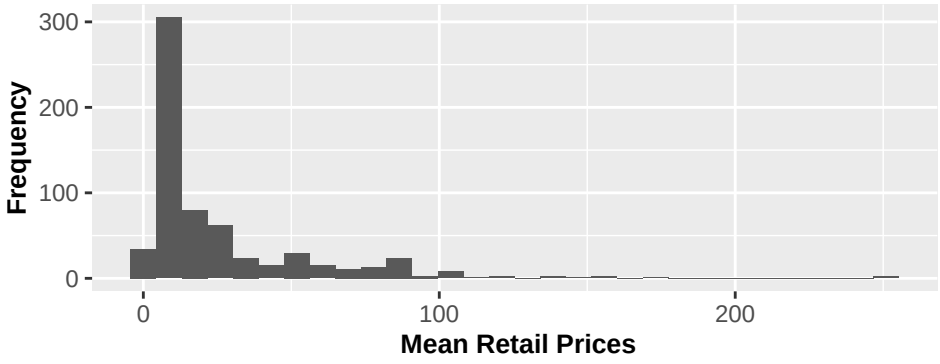
Figures mentioned in the above CIs are rounded to two decimal places and are based on heteroskedastic robust standard errors.

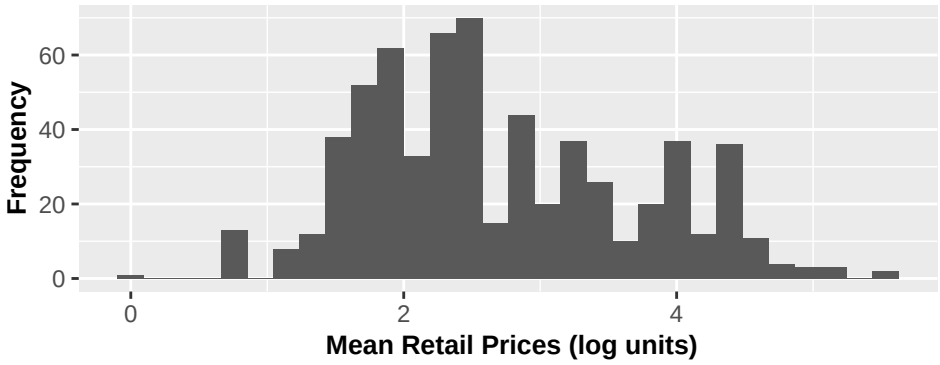
Here the beta coefficient is significant for the second spline of 'ln_online_price' estimating that on average the PED is between -1.66 and -0.37 when the price of the product on the website is greater than or equal to 2 log units, i.e around 7.4 Euros and we can reject the original null hypothesis with 99% confidence, i.e. on average PED is less than 0 in the population, holding law of demand true, keeping everything else constant.

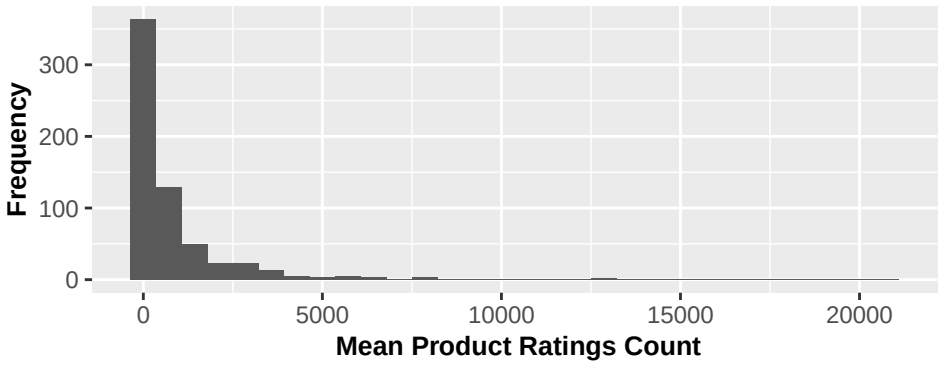
Distributions of important variables

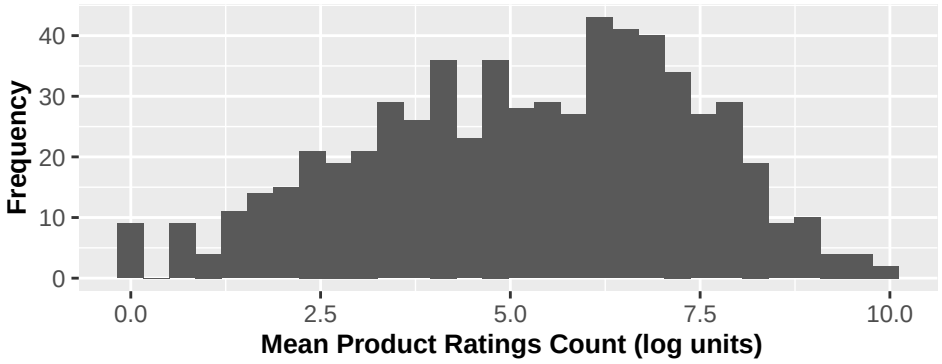



Frequency
60
40
20
0
0
5
10
15
Mean Merchant Rating Count (log units)

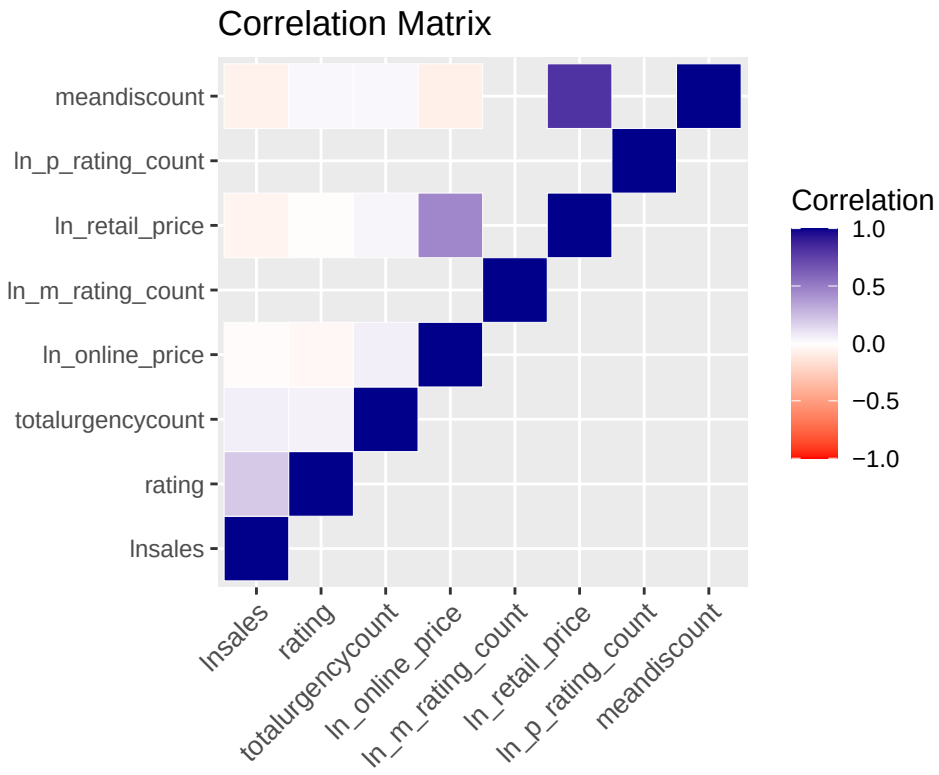

Frequency
300
200
100
0
0
100
200
Mean Retail Prices


Frequency
60
40
20
0
0
2
4
Mean Retail Prices (log units)


Frequency
300
200
100
0
0
5000
10000
15000
20000
Mean Product Ratings Count



Correlation Matrix



Original variable names have been used in this chart for accuracy purposes